Evidence-based Approach to Interacting with Open Student Models

Diego Zapata-Rivera, Eric Hansen, Valerie J. Shute, Jody S. Underwood, Malcolm Bauer, *Educational Testing Service,Rosedale Road. Princeton, NJ 08541 USA*
{DZapata, EHansen, VShute, JUnderwood, MBauer}@ets.org
www.ets.org

Abstract. Research efforts focused on developing "active reports" are currently underway. Active reports are designed to foster communication among teachers, students, and parents by listening to all stakeholders, using assessment information to guide teaching and learning, and to reconcile potential conflicts. Open student models can handle different views of the student. These views support different kinds of assessment information coming from a variety of sources (e.g. results from summative and formative assessments). As teachers, students, and parents interact with open student models, assessment claims (e.g. students' self-assessment, systems' and teachers' assessment of student knowledge) may evolve from unsupported claims to evidence-based arguments. Paths on this evidentiary argument space are characterized by rich interactions among sources of evidence that may challenge personal claims and improve metacognitive skills. This paper (a) presents an evidence-based approach to interacting with open student models that builds on existing research on open student models, evidentiary arguments and evidence-centered design, (b) illustrates this approach in the context of an assessment-based learning environment called Math Intervention Module (MIM), and (c) reports on a study carried out with Algebra teachers/assessment specialists aimed at evaluating the feasibility of implementing this approach in real settings.

Keywords. Evidence-centered design, open evidence-based argument structures, assessment-based learning environments, open student models, evidence-based interaction

INTRODUCTION

Advanced learning environments maintain information about students in the form of student models. These models can include cognitive and noncognitive aspects of the student. Information to populate the student model can come from a variety of sources. Standardized assessment programs, for example, are often accepted as sources of valid and highly reliable assessment information. However, they do not always provide information in a timely and detailed manner that can be used to support ongoing instruction and learning.

Not all sources of student model information share the same validity, reliability, accessibility and granularity parameters. Assessment information can be obtained from sources such as classroom quizzes, group activities, and self- or negotiated assessment activities. Each source of assessment information provides just a piece of the assessment puzzle. Integrating and making sense of diverse pieces of assessment information is a challenging activity that is often left as the sole responsibility of teachers. We are interested in exploring the use of computer-based tools that help teachers make appropriate use of assessment information from a variety of sources to enhance student learning.

Open student models (OSM) consider teachers, students, and sometimes parents to be more than just consumers of assessment information. In OSM, these participants play an active role by observing, updating, and acting based upon student model assessment information. Several researchers have explored the benefits and limitations of OSM.

Representational and interaction issues of OSM have been two of the main research areas explored. Kay (1995), for example, used a hierarchy of topics (partial models) in order to allow easy inspection by the user. Bull and Pain (1995) found that students seem to understand textually presented models. Bull and Nghiem (2002) showed that simple graphical or tabular representations of the model can be used to give learners a general idea of their strengths and weaknesses. In a more recent paper, Bull et al. (2005) reported that children, university students and instructors understood and used a variety of student model external representations. However, they also warn of possible negative effects when low-performance students explore student models of more capable students. Some of these students reported a negative effect on their motivation level and esteem.

Interaction with OSM can be done individually or collaboratively. It could include teachers, peers, parents or even artificial guiding agents. Different modes of interaction with OSM have been linked to various degrees of student reflection (Zapata-Rivera & Greer, 2002). Negotiated assessment using OSM (Bull, 1997; Brna et al., 1999) has been used as a mechanism to address conflicts resulting from discrepancies among different views of the student model and to facilitate formative dialogue between the teacher and the student. This dialogue process may result in a more accurate assessment and an optimized student model. Dimitrova (2003) describes an interactive open learning modeling approach in which the student and the system engage in interactive diagnosis. This process is aimed at supporting student reflection and improving the quality of the student model. Understanding of natural language is facilitated by the use of dialogue games, conceptual graphs, and sentence openers.

A factor that can potentially hinder OSM implementation in the classroom relates to teachers' perceptions of the fit between OSMs and current classroom practices. Some examples of such concerns include the following: (a) typical OSMs (e.g. symmetrical OSMs that consider teachers, students, and the system as equally important when integrating evidence of student proficiency) downplay the important role that teachers play in the classroom, (b) fear of giving students greater control over assessment (e.g. students could try to game the system), and (c) the system is sometimes seen as a complex, obscure, external entity that teachers do not understand and/or trust particularly when the system provides estimates of student knowledge and performance. These very real concerns of educators are explicitly addressed in this paper.

This paper presents an evidence-based approach for interacting with OSM. We call it an evidence-based interaction with OSM (EI-OSM). It has been inspired by our work on Evidence-Centered Design (ECD) of assessments, existing work on open student models, and ongoing research on creating assessment environments that communicate relevant assessment information to teachers, students and parents clearly (Zapata-Rivera et al., 2005). EI-OSM builds on Toulmin's (1958) work on argumentation, Schum's evidence based arguments (Schum, 1994), application of evidentiary reasoning in educational assessment (Mislevy, Steinberg & Almond, 2003), and reasoning about assessments given under nonstandard conditions (such as accommodated assessment administrations for individuals with disabilities) (Hansen, Mislevy, Steinberg, Lee, & Forer, 2005).

EI-OSM opens up the student model to teachers, students, and parents through the use of evidence-based argument structures that are used to organize pieces of evidence obtained from a variety of sources. Although teachers, students, and parents can interact with evidence-based argument structures using EI-OSM and the system offers a mechanism to reconcile various points of view, in this implementation of EI-OSM, the teachers are the ones who have the final word regarding which argument should be assigned the highest strength value. Unlike symmetrical OSM approaches, we provide an asymmetrical OSM in which teachers are given the possibility to override decisions based on available evidence. We argue that this particular aspect of EI-OSM, if implemented appropriately (i.e. proper support, tools and educational materials should be available to facilitate teacher interaction with the system and avoid cognitive overload), could be critical when considering using this system in real settings. This teacher-oriented approach is consistent with approaches used by successfully deployed intelligent tutoring systems (e.g. Koedinger, Anderson, Hadley, & Mark, 1997).

This paper presents the ECD framework, describes Active Reports (ARs) and EI-OSM and explains how they have been used to open up ECD-based student models, illustrates the EI-OSM approach in the context of an ETS project called the Math Intervention Module (MIM) (Kuntz et al., 2005; Shute, in press), reports on a study carried out with eight Algebra teachers/assessment professionals aimed at evaluating the feasibility of implementing this approach in real settings, discusses our findings, presents related work, and concludes with a discussion of the potential of EI-OSM and our future work.

In order to implement EI-OSM, it is necessary to rely on an evidence-based framework that facilitates the creation of open evidence-based argument structures. This framework is provided by Evidence-Centered Design. The next section describes evidence-centered design and its three main models (i.e. proficiency/student, evidence and task models).

EVIDENCE-CENTERED DESIGN

Evidence-Centered Design (ECD) (e.g. Mislevy, Steinberg & Almond, 2000, 2003) is a methodology for assessment design employed at Educational Testing Service that emphasizes a logical and explicit representation of an evidence-based chain of reasoning, with the goal of ensuring the validity of assessment results. This evidence-based chain of reasoning is articulated through the development of proficiency, evidence, and task models, and by starting the process with a Prospective Score Report (PSR). Each step in the design process is logically linked to the previous and following steps, aligning all the steps in the development process. ECD is intended, among other things, to support the development of innovative tasks, while maintaining or enhancing the validity of assessment results. The focus on early consideration of reports (through PSR) has led to more useful score reports. ECD helps us understand what to measure and how to explain what is being measured. Figure 1 depicts the three central models of ECD - the proficiency, evidence, and task models.

To design an assessment, the purpose and claims must first be defined. The PSR is a tool that helps (a) establish the claims to be made about a test-taker who passes or fails the test, (b) define the recipients of the score reports (i.e. what to report to whom), and (c) what needs to be measured in order to satisfy the claims. The PSR should be used at the beginning of assessment design, with the reporting measures continually being refined, as necessary, through the remainder of the assessment design process. By doing this, the assessment is developed to support the scores identified on the PSR. This is done while keeping constraints (e.g. test administration time limits) in mind.

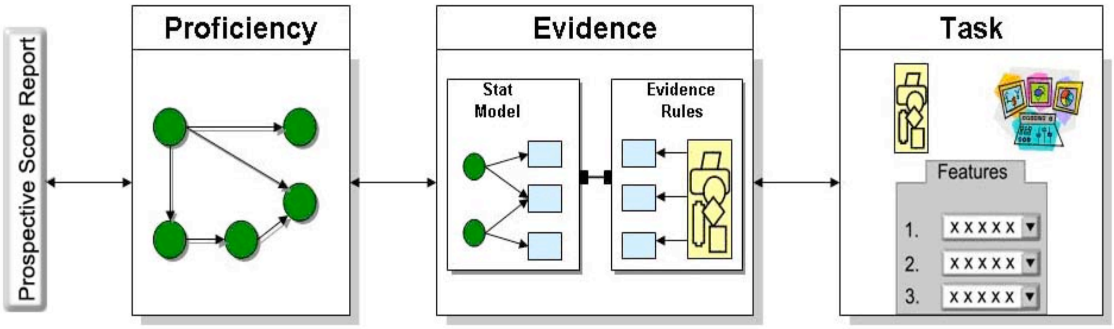


Fig.1. ECD central models.

Proficiency models, also called *student models*, further define the Knowledge, Skills, and Abilities (KSAs) that need to be measured and reported. These KSAs are clearly linked to claims about the test taker (*assessment claims*), and are defined to support the scores or subscores that are defined in the PSR. It is during this phase that the characteristics of the test taker are considered (e.g. age range, special needs) so that the claims can be supported for diverse individuals within the target population.

Evidence models define what observables (observations) will provide the evidence required to support the claims defined in the proficiency model. Characteristics such as knowledge, intelligence, aptitude, attitudes, etc., cannot be ascertained directly but must be inferred based on observables, which have been extracted from the work products produced as part of the task model. In some cases, observables are based only on final work products (e.g. the solution to a mathematics problem) and in other cases may be based on intermediate products (e.g. partial or intermediate steps in solving a mathematics problem). Evidence models need clear links to the KSAs in the proficiency model. A key job of the ECD assessment designer is to define tasks (see below) that will evidence KSAs in the proficiency model. For each piece of evidence, the ideal behavior must be defined, as well as departures from that behavior that may provide alternate forms of evidence.

The evidence model has been divided into two main components: a scoring component (*evidence rules*, in Figure 1) and a measurement component (*statistical sub-model*, in Figure 1). Evidence rules are used to generate observables based on identification of salient features from work products. Evidence rules typically involve the use of keys, rubrics, and scoring rules. On the other hand, statistical sub-models are used to represent relationships between observables and the KSAs in the proficiency model (e.g. how pieces of evidence are accumulated). In a probabilistic framework, for example, statistical models might take the form of IRT (item response theory) or Bayesian models. Claims made in operational use of an assessment are supported by the same chain of reasoning that guides test development, but in the opposite direction - from tasks, to evidence and finally to the claims that will be reported. Such assessment claims are reported in terms of quantitative or category scores (e.g. pass-fail, basic-proficient-advanced).

Task models describe the assessment tasks, including a specification of the work products (e.g. record of a mouse click, essay, etc.) to be produced by the student. The range of permissible values or settings for a task feature may be specified, thus constituting a model of a class or family of tasks. Creating high-level task models helps avoid premature decisions about task or item types (e.g. should they be multiple choice or short answer?). Logically, the models would be defined before task definitions, or shells, but often they are made after a few task shells have been defined. Appropriate task models capture all salient work product features that have been specified. Task models are linked to a proficiency model via evidence models. These task model designs must take into consideration the characteristics that will affect validity and difficulty, as well as constraints such as time, cost, scoring capacity, and reporting time. If the task models cannot meet all these constraints, the claims must be reassessed. Similarly, defining the tasks may identify irrelevant KSAs, unnecessarily difficult language, or other test fairness problems that need to be changed. Task models describe the types of stimuli that might be used, and describe in general terms what the test taker will be asked to do. They describe the task elements that must be present, as well as those that may vary from task to task.

ECD is not limited to a particular cognitive domain model, type of task, type of evidence, model representation scheme, or scoring model. Instead, ECD provides general and flexible principles and tools to guide and support the assessment design process. ECD helps build in validity, provides clear documentation, and it helps keep focus during the assessment design process.

OPENING ECD MODELS

Evidentiary chains of reasoning created by using ECD have been externalized in the form of open evidence-based argument structures (EI-OSM) following the ideas expressed in the active reports (ARs) approach.

Active reports and the formative loop

The Active Reports (AR) approach (Zapata-Rivera et al., 2005) describes a conceptual framework for opening ECD models. In this approach, evidence of student performance from various sources (e.g. standardized assessments, classroom assessments, teachers, students, and parents) is integrated and used to support student learning. ECD models hold valuable student assessment information that can be made available to teachers, students, and parents through the use of ARs.

ARs are active entities that can be used to convey useful information, support communication among students, teachers, and parents and at the same time serve as a mechanism to acquire additional evidence of student knowledge and performance and use it to update the student model. ARs can present different views of the student model (e.g. an AR view for the teacher), as well as different aspects of the student model (e.g. assessment claims, supporting evidence, different levels of granularity, and different external representations).

Assessment claims and supporting evidence gathered from teachers, students, and parents can be linked to the ECD framework as new evidence that corroborates (or fails to corroborate) an existing assessment claim or provides support for a new, alternative explanation. Using ARs, a view of the ECD models containing information about student knowledge and supporting evidence (including tasks) is presented to these stakeholders, who interact with it and provide additional insights that are gathered during the learning process and can be incorporated into the assessment argument. We argue that by listening to teachers, students, and parents, ECD models are enhanced and become instrumental in supporting a formative learning process based on a view of the student model that changes in light of new evidence.

Figure 2 depicts the AR framework. After the student has taken a test (i.e. *Testing Event*), an AR instantiation is generated and presented to the student, teachers, and possibly parents. Assessment claims and supporting evidence gathered through ARs are modeled as a source of evidence that is captured in the ECD framework as part of the evidence model (see double arrows). ARs can include information such as assessment claims based on KSA levels (i.e. proficiency model information), and supporting evidence used for backing up assessment claims (i.e. evidence and task model information). Assessment information from prior testing events (e.g. past performance) is available for inspection (see dashed line) and maybe used to initialize the ECD models. The presentation model is used to select and present appropriate tasks to the student based on the student model (e.g. adaptive sequencing of tasks in a computer adaptive testing [CAT] environment).

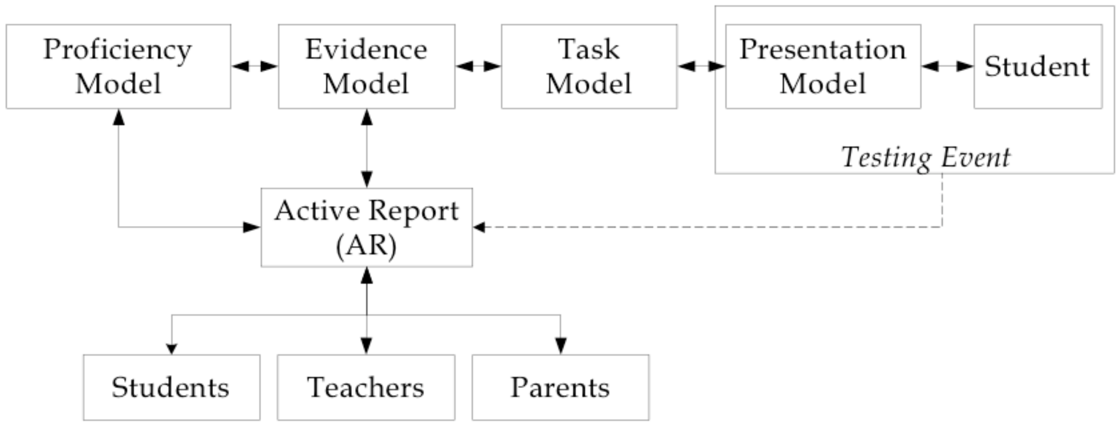


Fig.2. AR framework. ARs are used to present evidence of student knowledge/performance to stakeholders, and gather additional evidence from stakeholders.

Interacting with ECD models through ARs serves different purposes: (a) learning can be focused on topics that the student needs to learn or is ready to learn, (b) student reflection can be supported by exposing the student to evidence of his/her performance and how it is used to support assessment claims, and (c) teachers and parents can also interact with this information and use it to facilitate teacher-parent communication and to provide appropriate and timely feedback to students. Thus, as a communication tool, ARs can support synchronous and asynchronous formative communication among teachers, students, and parents based on the student model. Results from an initial assessment can be used to support a learning process based on continuous-, self-, collaborative-, and negotiated-assessments.

ARs can be implemented as interactive evidence-based argument structures that connect proficiencies to outcomes and allow for representing alternative explanations and supporting evidence originating from various stakeholders. The next section presents this approach.

Evidence-based interaction with OSM (EI-OSM)

EI-OSM extends earlier work in OSM by focusing on evidentiary arguments and goes beyond other evidence-based work by providing explicit mechanisms for teachers, students, and others to negotiate the substance and meaning of the assessment argument. EI-OSM provides teachers with an evidence-based framework that facilitates understanding and integration of assessment sources. This approach may have two benefits for student learning: (1) provide estimates of proficiency that are more valid for certain intended purposes, thus helping to guide teachers' instructional efforts, and (2) build students' critical thinking and reflection skills.

In a broad sense, evidence refers to anything that can be used to "prove" or "disprove" something. It can include people's testimonies, facts, or physical objects that point to a particular direction or suggest the occurrence of an event. It also includes arguments properly derived from existing evidence through chains of reasoning. Handling of evidence in ECD is facilitated by the use of chains of reasoning that connect the student's performance on a particular test to assessment claims about his/her KSAs. EI-OSM opens-up ECD evidence-based argument structures that support each assessment claim.

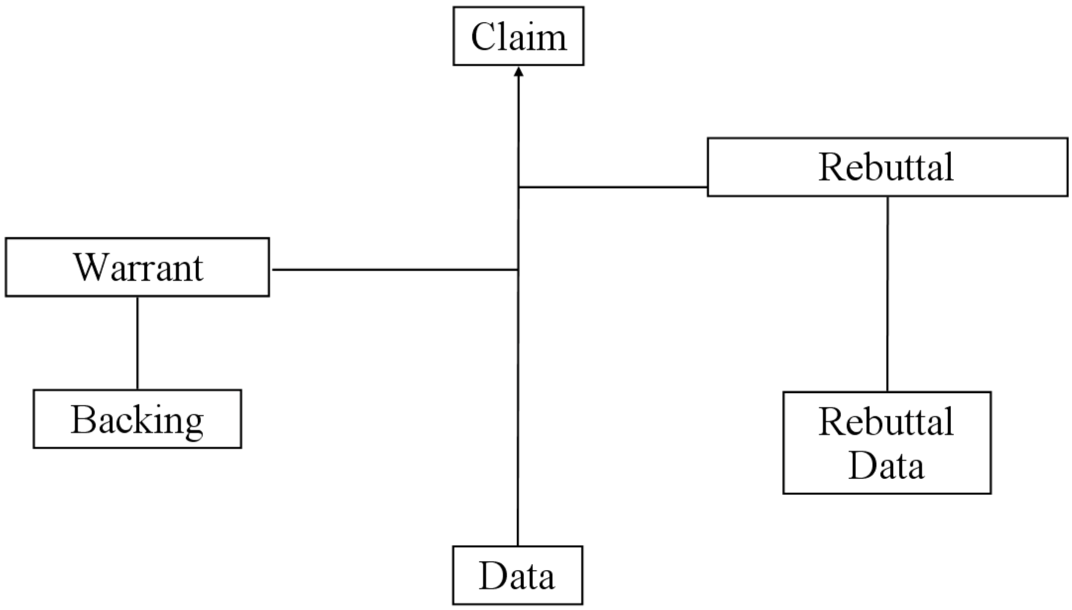


Fig.3. Argument structure.

Figure 3 shows the structure of an argument, adapted from Hansen, Mislevy, Steinberg, Lee, and Forer (2005), based on Toulmin (1958). Toulmin's work on argument structure has been used in a variety of research areas including collaborative argumentation (e.g. Suthers, Weiner, Connelly, & Paolucci, 1995; Jeong & Juong, 2007), dialogue games (e.g. Bench-Capon, 1998), knowledge engineering (e.g. Carbogim, Robertson, & Lee, 2000), and external representations (e.g. Rouane, Frasson, & Kaltenbach, 2002). EI-OSM uses a simplified version of Toulmin's argument structure to externalize, organize and evaluate assessment claims and supporting evidence.

Following are explanations of key elements of Toulmin's argument structure. The *claim* is the assertion that is the subject of the argument. In the context of an *assessment* argument, it may be an assertion about what a student knows and can do, for example, a fictitious student named Clarissa has a *low* (as opposed to high) level in the targeted proficiency in Algebra I content. The *data* (observable) is information that supports the claim. For example, in an assessment argument, key examples of data might be a student score (for an item [question] or the whole assessment). The *warrant* is the generalization that licenses the inference from data to claim. An example of such a generalization would be that, "If a person has low Algebra I proficiency, they will obtain low scores and if they have high Algebra I proficiency they will obtain high scores." The *backing* is information - obtained through theory and experience - that supports or backs the warrant. The rebuttal (which might be referred to as an "alternative explanation") is an assertion that seeks to defeat the basic argument (the claim, data, and warrant). An example of a rebuttal would be that Clarissa actually possesses a *high* level of Algebra I proficiency; and *rebuttal data* supporting the rebuttal might be, for example, that she performed poorly due to receiving the Algebra I test in a format (regular sized font) as opposed the large font that she needs because of a visual impairment.

It may be useful to make more explicit the different kinds of rebuttal of the basic (claim, data, warrant) argument. One can argue, for example, against the data ("The low score resulted from an incorrect key"), or against the warrant ("The generalization that 'If a person has low Algebra I proficiency, they will obtain low scores and if they have high Algebra I proficiency they will obtain high scores' is false"),[1] or against the claim ("Clarissa actually has high ability").

Our evidence-based approach for OSM (EI-OSM) allows teachers, students, and others to interact with a rich representation of the argument. Consider the following possible capabilities:

1. *System display of key parts of the argument*. The system might allow the student to not only examine the claims (the system's assertions about student proficiency), but also to examine the data, warrants, and rebuttals.
2. *Provisions for students to challenge particular aspects of an argument*. The student might be allowed to challenge the argument. In addition, the system might structure the interaction to elicit from the student the specific nature of the challenge. For example, the student might be allowed to challenge the claim, the data, the warrant, or the rebuttal (as suggested above). The student might be allowed to select one or more specific objections, perhaps with accompanying rationales. The opportunity to write a textual response or to attach supporting documents might be provided.
3. *Provisions for students to propose a different argument*. Instead of only responding to the argument of the teacher or system, the system might allow students to propose an alternative argument. One can imagine an argument creation wizard that would walk students through the creation of argument structures, guiding them in the articulation of claims, warrants, and data.
4. *Methods for negotiating the meaning of argument elements*. One can imagine a system-mediated process for establishing the information as facts that can then be incorporated into arguments. For example, in response to the argument set forth by the teacher (through the system), the student might propose that her high score on an earlier assessment or class assignment be considered as defeating the teacher's (or system's) assertion that the student has low proficiency in a certain Algebra I topic. The system might allow the student to propose the earlier test score as a fact; the teacher might then be alerted to verify that the student did indeed obtain the score. (Of course, the system could contain a wide range of scores which the teacher has already accepted as facts.) An argument or portion of an argument (e.g. claim, data, warrant) constructed using such "facts" might be considered as having greater strength

[1] Verheij (2001) identifies additional specific classes of rebuttal (p. 9).

(or credibility) than an argument constructed with information that had not been approved or vetted as facts.

Students and teachers interact with open evidence-based argument structures to refine assessment arguments, explore existing evidence, and facilitate instructional dialogue. Various guiding mechanisms can also be implemented to support interaction with open evidence-based argument structures (e.g. guiding protocols, collaborative tools, artificial guiding agents, negotiation with the teacher, and computer-supported dialogue). Guided interaction with open Bayesian student models has shown the presence of student reflection, which also can contribute to student learning (Zapata-Rivera & Greer, 2003).

Evidence-based interaction with OSM supports the refinement of assessment claims by organizing them into open evidence-based argument structures that can be explored, annotated, and assessed. Evidence-based argument structures connecting evidence of student performance to proficiency levels can be used to inform decision-making processes (e.g. informing teachers and parents, planning lessons and special instructional programs, guiding self-directed learning, and adapting materials and lessons as part of an adaptive learning environment). Because this approach supports the evolution of evidence-based argument structures, both the assessment-based learning environment and the educational stakeholders involved can benefit.

Sources of evidence can be classified based on parameters such as credibility, relevance and inference force (Schum, 1994). As new assessment claims and supporting evidence are added to an argument structure, new branches are created, which results in a tree structure of alternative arguments and corresponding supporting evidence. The strength of each piece of supporting evidence added by teachers or students to the system is estimated based on parameters such as its relevance (i.e. degree to which the evidence relates to a proficiency), credibility of the source (e.g. homework vs. in-class individual activities), date, and grading information. Initial relevance and credibility parameters can be provided by teachers or given to them (i.e. predefined packages of learning materials and activities). Strength of evidence offered by the system in support of a particular assessment-claim is defined as the marginal probability associated to a particular state of the proficiency (e.g. Prob (Proficiency_i = *Low* | *available evidence*). Using ECD, different statistical models can be employed to connect task performance to proficiencies (e.g. item response theory or Bayesian networks). These statistical models take into account the number and difficulty of the tasks and how they connect to the proficiency model (e.g. Q-Matrix information (Tatsuoka, 1995)).

Weights are used to combine evidence parameters into a single measure of strength (both for single and multiple pieces of evidence in support of a single assessment claim). A similar approach is used to integrate pieces of evidence from various sources. Preliminary work on inspecting Bayesian student models (Zapata-Rivera & Greer, 2004) shows how the student's, teacher's and system's views of the student model can be integrated using Bayesian networks.

Teachers and students use a graphical tool for creating open evidence-based argument structures (i.e. add new arguments and supporting evidence). Teachers can also select and attach pieces of evidence to proficiencies (i.e. map tasks to proficiencies). Students can choose supporting evidence from an existing list of available pieces of evidence (previously defined by teachers) or add their own (e.g. additional explanations, and relevant pieces of evidence not included in the system). Thus, EI-OSM supports the use of different kinds of supporting evidence, which can be used to change the strength of an argument in a variety of ways. Supporting evidence that has passed the inspection of the teacher, for example, is considered stronger than an unsupported assessment claim by the student.

The next section presents MIM, an assessment-based learning environment that was built following ECD principles. We use MIM to demonstrate the main features of EI-OSM.

MATH INTERVENTION MODULE (MIM)

MIM (Kuntz, et al., 2005; Shute, in press) is an online prototype designed to help students become proficient in applicable mathematics standards. The initial focus is on Algebra I. The module is based on a proficiency model that describes the skills that must be mastered to be judged proficient at a standard. Each module presents students with open-ended questions dealing with the various skills identified in the proficiency model. These questions require the student to respond with (a) a number, (b) an expression or an equation, (c) a graph, and (d) text, all of which are automatically scored.

MIM is an assessment-based learning environment. Assessment-based learning environments make use of assessment information to guide instruction. Some examples of assessment-based learning environments include: Web-based cognitive tutors called *assistments*, the merging together of assessment with instructional assistance into one system (Razzaq et al., 2005), SIETTE (Conejo et al., 2004), LeActiveMath system - xLM (Morales, Van Labeke, & Brna, 2006), English ABLE (Zapata-Rivera et al., 2007), and ACED (Shute, Hansen, & Almond, 2007).

Diagnostic Feedback. All responses in the MIM system are automatically evaluated, with immediate feedback provided to the student. Feedback is directed at the error that the student has made, and is not simply, "Wrong. Please try again." Similar to a human tutor, MIM attempts to give some indication of why the student's answer was wrong. The student is given three attempts to answer each question correctly, with progressively more detailed feedback provided along the way. The correct answer, with an associated rationale, is presented if the student answers incorrectly three times. In addition, if the student is judged to be in need, the module presents a short (i.e. 2-4 minute) instructional video that covers the problematic concept or skill. These "instructional objects" reinforce the learning that is taking place as the student works through the questions and reads the feedback.

Instructional Objects. A specific instructional object (IO) is presented when a student requires all three levels of feedback. There are currently about 16 IOs produced for the current MIM prototype. Within an IO, the flow of instruction proceeds as follows: (a) introduce the topic using concrete and engaging context, (b) state a particular problem that needs solving, (c) provide relevant definitions, (d) illustrate the concept within different examples (both prototypical and counter-examples), (e) provide practice opportunities and interactivity, and (f) conclude with summary and reflection screens.

Practice Opportunities. The teacher has the option of assigning multiple-choice questions for additional practice on each skill. The teacher can (a) require these practice questions of all students who seem not to have mastered the skill, (b) make the practice questions optional, or (c) configure the module so that the practice questions are not delivered.

Integrating Knowledge and Skills. The final section of each intervention module is a set of integrated open-ended questions that deal with a common theme or contextual situation. These questions reflect the standard as a whole. Like the open-ended questions earlier in the module, these integrated questions involve responses that require the entry of a number, an expression or an equation, a graph, or text.

ECD Models in MIM

Proficiency Model. As described earlier, a proficiency model may include KSAs that are of interest for the assessment or learning environment. In MIM, the proficiency model includes the skills that must be mastered to be judged "proficient" in relation to a specific standard, and displays the relationships between these skills. The initial MIM prototype uses a proficiency model that analyzes the standard, "Translate word expressions to symbolic expressions or equations and then solve and/or graph" (see Figure 4). By working down the model, one can see how the component skills are isolated. The current version of MIM implements eight skills (grey nodes in Figure 4).

The various elements of MIM - open-ended questions, instructional videos, and multiple-choice practice questions - are presented to the student according to a carefully planned instructional design. We used ECD to develop the underlying proficiency model, scoring rules, and informative assessment tasks, and incorporated research-based features into MIM to support learning (e.g. timely diagnostic feedback, tailored content, and multiple representations of concepts).

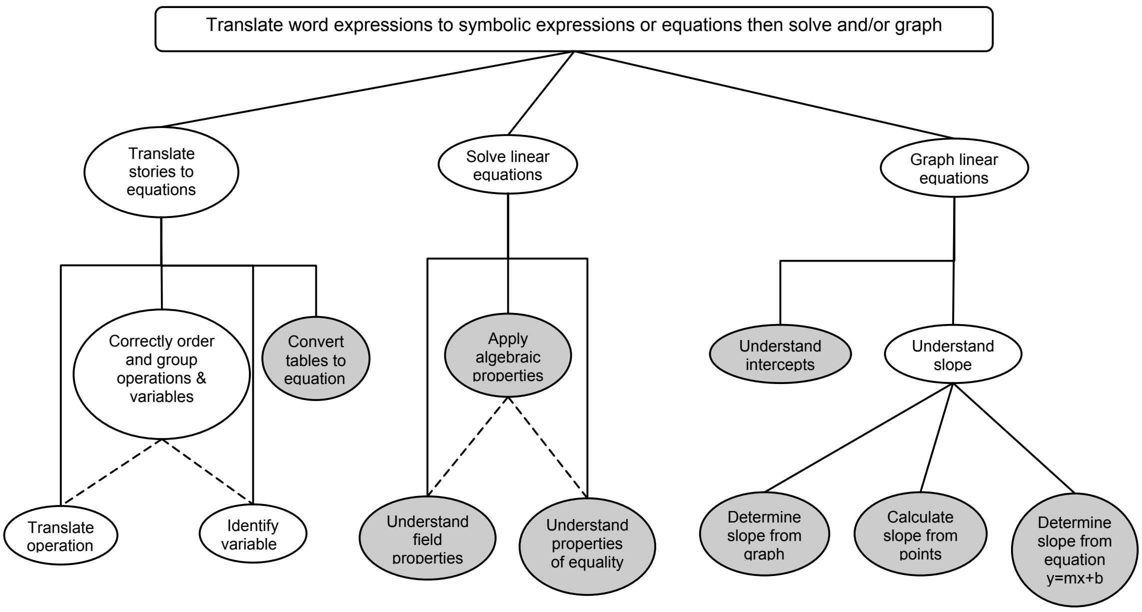


Fig.4. Proficiency/Student Model for the MIM prototype.

Evidence and Task Models. Evidence of student knowledge or ability comes from "easy" and "hard" tasks. Evidence gathered using these tasks can be linked to a single external representation or assessment aspect of a skill (e.g. numerical aspect) or to a group of them (i.e. "integrated tasks"). Figure 5 shows an example of how elements of proficiency, evidence and tasks models are connected in MIM. As students interact with MIM's tasks, evidence of student knowledge is propagated throughout these models. The proficiency model maintains the current knowledge profile of each student. Assessment claims based on ECD models are supported by student performance on particular tasks and by information on how tasks relate to skills.

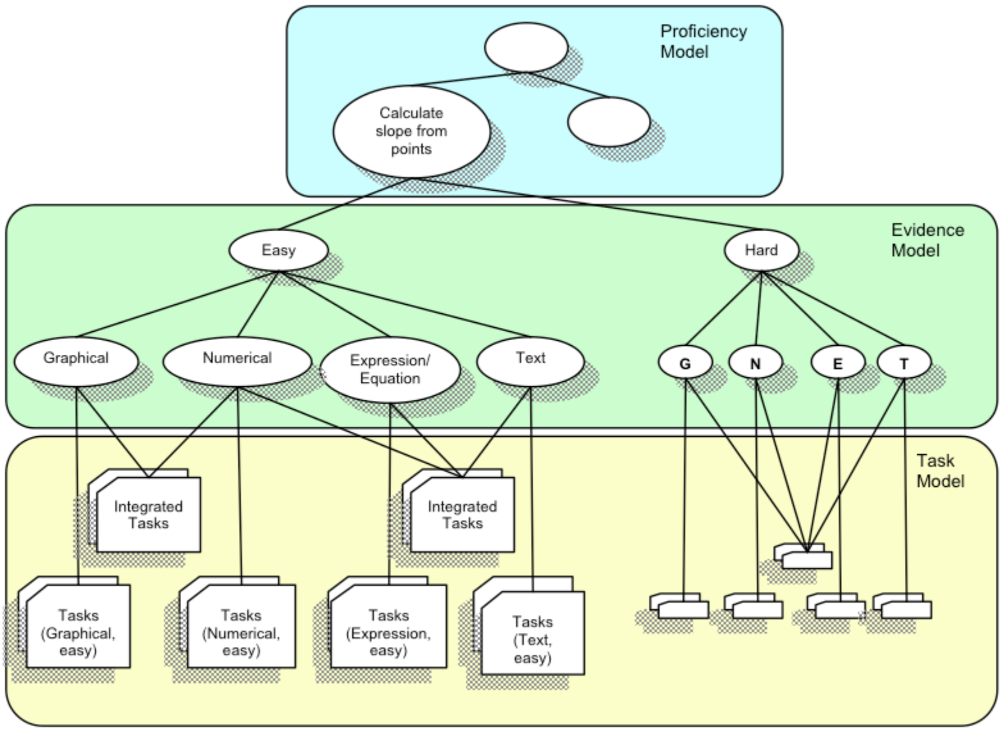


Fig.5. Fragment of ECD Models in MIM.

Exploring Evidence-based Argument Structures

A proficiency map (see Figure 6) is used to present an *aggregate* view of the proficiency/student model (i.e. status of the student model after having integrated various pieces of evidence from different sources). Color has been used to depict proficiency levels (e.g. green = high proficient level - "*Go ahead,*" yellow = medium proficient level - "*Warning*," and red = low proficient level - "*Stop*"). Figure 6 has been annotated using *G=green, Y=yellow, O=orange, and R=red* to make it easier to understand in black and white printing. This proficiency map has been generated using ViSMod (Zapata-Rivera & Greer, 2002; 2004).

By clicking on a proficiency node, teachers, students or parents can see the evidence-based argument structure (i.e. assessment claims and supporting evidence) attached to that particular node for a particular student. Figures 7 and 8 show MIM's assessment claim and supporting evidence regarding Clarissa's knowledge of "*Calculate slope from points.*" The assessment claim with the highest strength value appears at the top of the screen (at this point only MIM's assessment claim is available). Supporting evidence for assessment claims is placed on either side of the argument backbone line (see Figure 9).

Assessment claims in MIM are assertions regarding what a student knows and can do based on the state of the proficiency model and properly backed up by evidence and task model information. For example, *Clarissa has a low level in* "*Calculate slope from points,*" given that she has answered several questions covering several evidence model aspects of the skill incorrectly. EI-OSM provides a way for teachers, students, and parents to initiate a formative interchange based on evidence-based arguments. These arguments get refined as more evidence is provided. This will influence the final or aggregate view of the proficiency model.

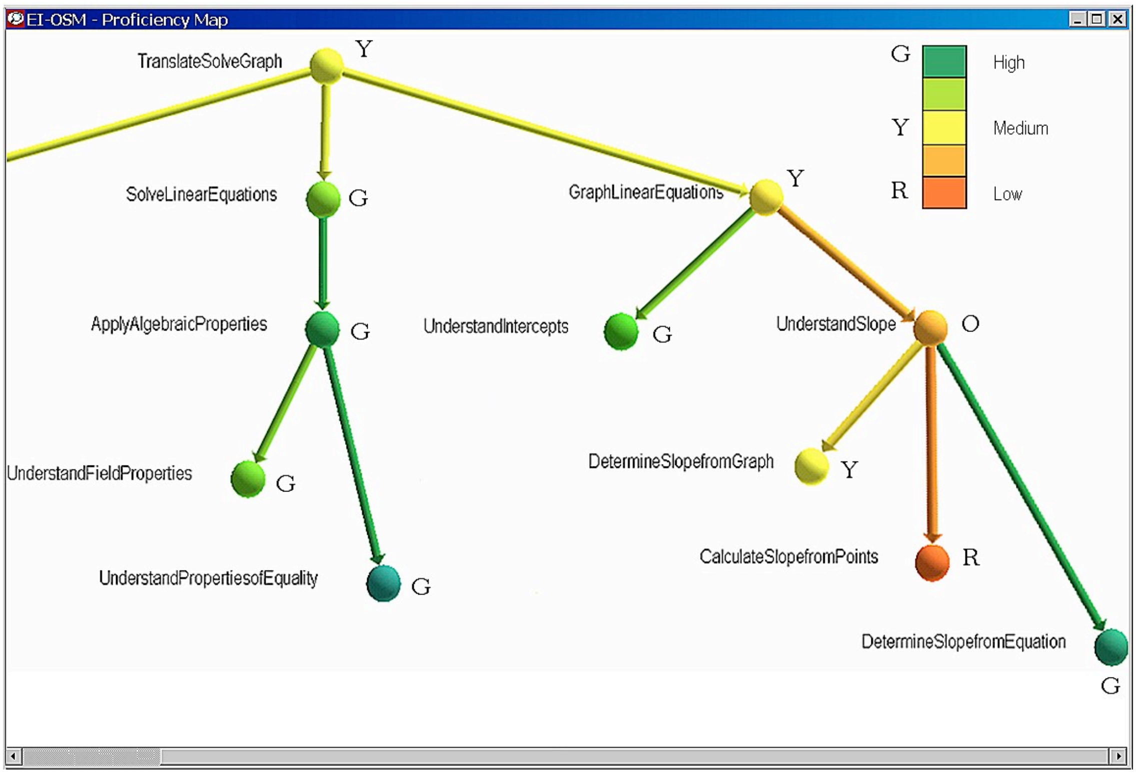


Fig.6. Exploring a Proficiency Map.

Supporting evidence for a particular assessment claim can come from direct and indirect paths of the ECD structure. For example, let us consider the following assessment claim: *Clarissa has a medium level in* "*Graph linear equations*" (see Figure 6). Evidence of student performance/ knowledge regarding this skill can be gathered directly from tasks attached to it as well as from tasks attached to skills in the subtree beneath it. In addition, evidence could also arrive indirectly from other parts of the proficiency model and then be propagated through a common parent (e.g. "Translate word expressions to symbolic expressions or equations and solve and/or graph").

Students can explore MIM's assessment claims and decide to offer an alternative explanation and supporting evidence. Following with our example, Clarissa feels that MIM's assessment claim is not accurate and decides to challenge it. She adds a new assessment claim (*My knowledge level of* "*Calculate slope from points*" *is* "*High*") and explains that she already knows how to solve those items. See Figure 9.

Clearly, there is a conflict that needs to be resolved in order to help guide Clarissa's learning and at the same time improve the accuracy of the student model. Figure 9 shows two indicators of strength - one for each side of the argument structure. The strength indicator works as a measure of the credibility, relevance, and quality of the supporting evidence provided. It is used to balance each side of the argument structure. In this case, MIM's argument strength is superior to that of Clarissa's alternative argument.

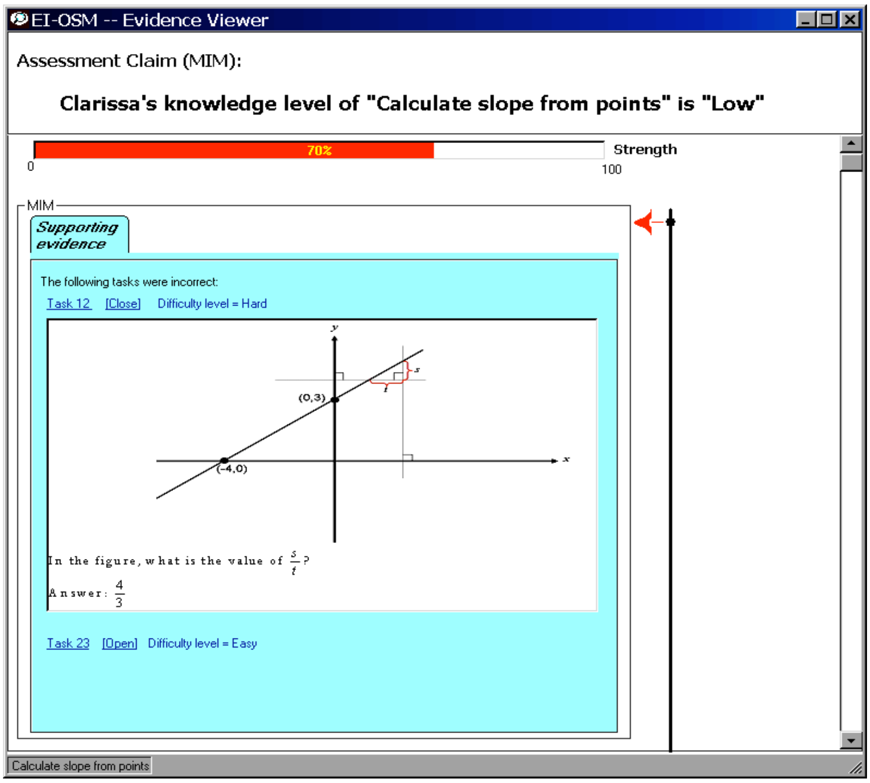


Fig.7. Supporting evidence for MIM's assessment claim - Task 12, a "hard" task linked to "Calculate slope from points" through numerical and graphical aspects of the evidence model.

By clicking on the supporting evidence tab (see Figure 9), Clarissa adds pieces of supporting evidence resulting in a stronger argument (see Figure 10). Students can select pieces of supporting evidence from a list of available pieces of evidence previously defined by the teacher or add their own (e.g. an explanation of how to solve the tasks as well as links to educational materials and other information that can help them demonstrate their knowledge). In this case, Clarissa has selected two pieces of supporting evidence from the list. As mentioned before, evidence parameters and weights are used to estimate a new strength value for Clarissa's argument. Relevance and credibility values are assigned by teachers or taken from predefined packages of educational materials. Date and grade information appears automatically as soon as this information becomes available. When new pieces of supporting evidence are added to the system, default parameters are assigned based on information elicited from the teacher in advance (e.g. default values assigned to predefined categories). Students classify new pieces of evidence as they enter them into the system based on a list of predefined categories. Teachers can change these values at any point in time.

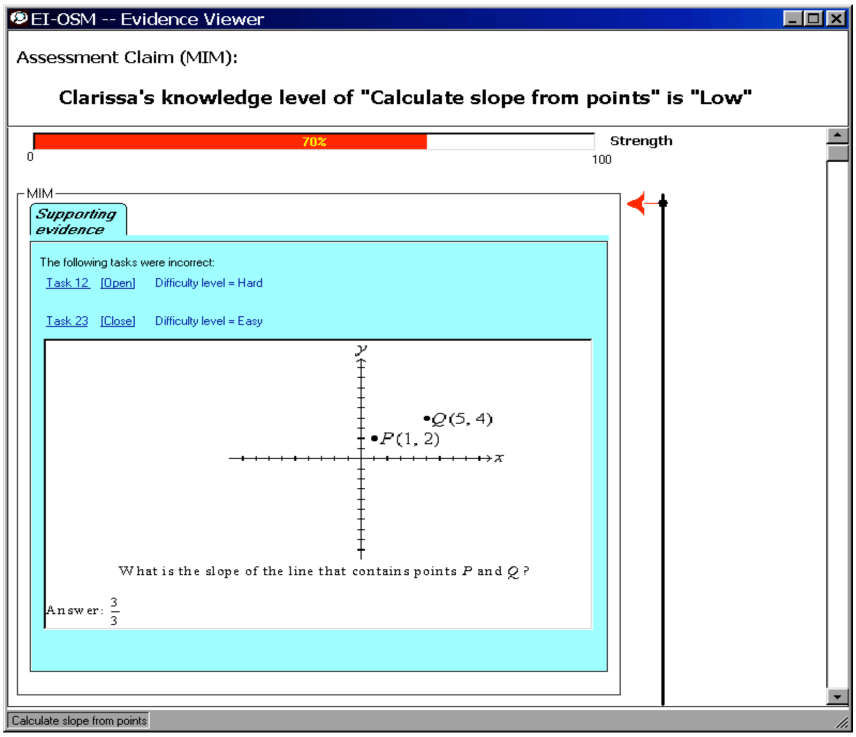


Fig.8. Supporting evidence for MIM's assessment claim - Task 23, an "easy" task linked to "Calculate slope from points" through numerical and graphical aspects of the evidence model.

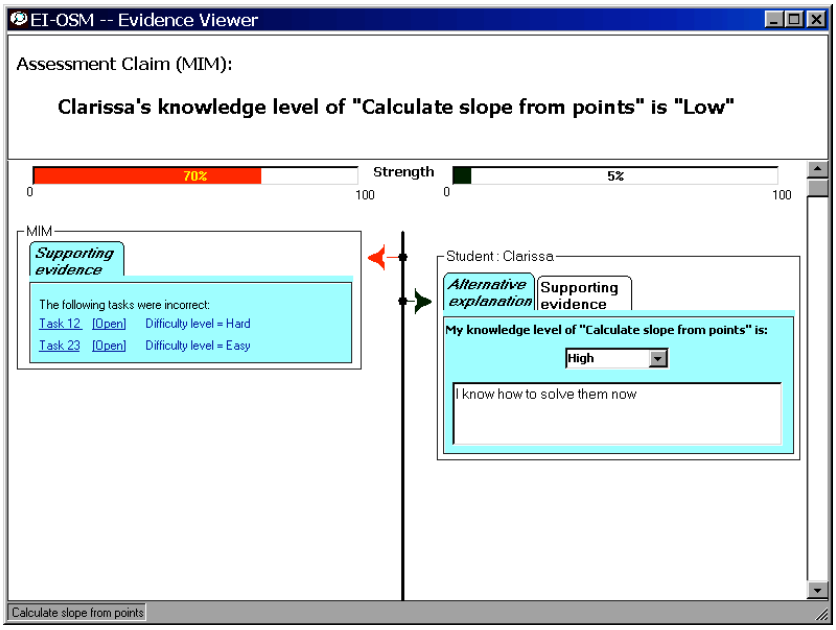


Fig.9. Clarissa offers an alternative explanation (no other supporting evidence than Clarissa's assertion "*I know how to solve them now*" has been provided at this point).

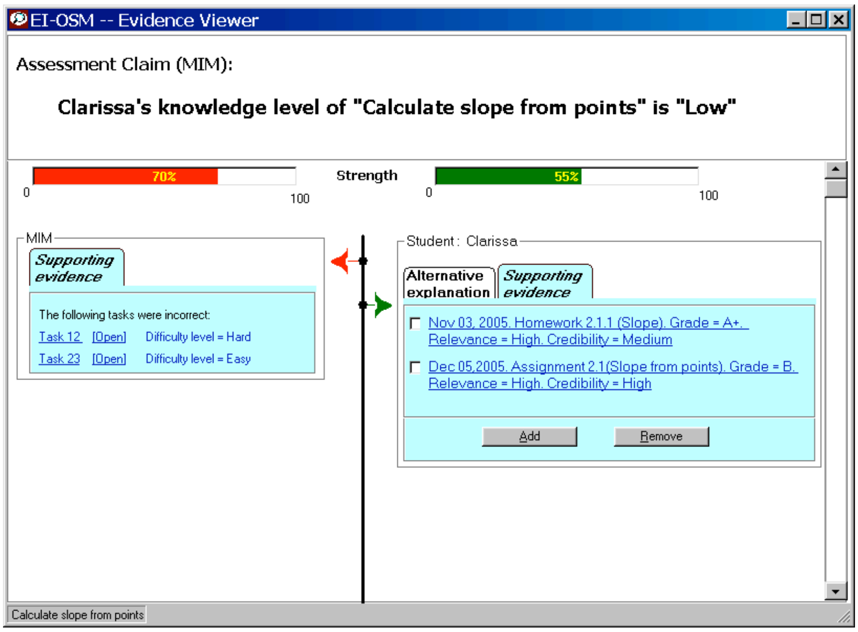


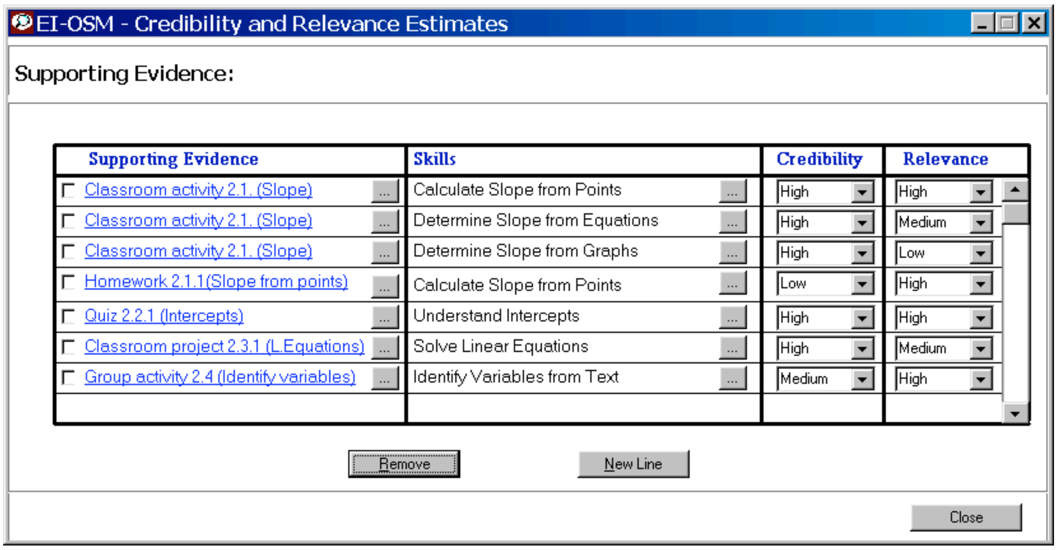
Fig.10. Clarissa adds several pieces of supporting evidence making her argument stronger.

As Clarissa's assessment claims get further refined, their strength is used to revise MIM's assessment claims. This is done by propagating Clarissa's new supporting evidence through the evidence model and subsequently to the proficiency model following the ARs approach presented earlier (see Figure 2).

Figure 11 shows one of several graphical interfaces used by teachers to add pieces of supporting evidence, attach them to proficiencies, and set credibility and relevance values. Other graphical interfaces make use of numerical values (e.g. 1, 2, 3) instead of labels (e.g. high, medium, low) when defining relevance and credibility levels. We have also explored assigning strength values to pieces of evidence directly (no relevance and credibility values are used in this case). As showed in Figure 11, a single piece of evidence (*Classroom activity 2.1. Slope*) can be assigned to several skills using different relevance values. Some other values (e.g. credibility values) are set by default using a list of assignments previously defined by teachers (e.g. default credibility value for homework = low).

At this point, Clarissa can be asked to provide additional supporting evidence (e.g. solve additional tasks on this skill using MIM). A list of appropriate items can be selected based on the current state of Clarissa's proficiency, evidence, and task models including items linked to unexplored aspects of the evidence model (e.g. items linked to the "*Text*" and "*Expression*" assessment aspects of the model, assuming that Clarissa has already solved tasks related to the "*Graphical*" and "*Numerical*" ones) as well as items with different degrees of difficulty. Solving these items correctly will increase the strength of Clarissa's argument and at the same time the results will be used to update Clarissa's student model in MIM. Solving additional items without a clear understanding of the kinds of items and the difficulty level expected by MIM's evidence model could result in unfruitful interactions, serving as a source of frustration. For instance, Clarissa may wonder why the strength of her argument remains "low" after having solved 3 more tasks, not realizing that the items solved were all "*easy*" and related to the same assessment aspect in the evidence model. Figure 12 shows a message used in EI-

OSM to provide adaptive feedback to students on tasks and educational materials, and clear up such potential confusion.



Supporting Evidence	Skills	Credibility	Relevance
Classroom activity 2.1. (Slope)	Calculate Slope from Points	High	High
Classroom activity 2.1. (Slope)	Determine Slope from Equations	High	Medium
Classroom activity 2.1. (Slope)	Determine Slope from Graphs	High	Low
Homework 2.1.1(Slope from points)	Calculate Slope from Points	Low	High
Quiz 2.2.1 (Intercepts)	Understand Intercepts	High	High
Classroom project 2.3.1 (L.Equations)	Solve Linear Equations	High	Medium
Group activity 2.4 (Identify variables)	Identify Variables from Text	Medium	High

Fig.11. Graphical interface used by teachers to add pieces of supporting evidence, attach them to proficiencies, and set credibility and relevance values.

Educational materials are offered based on the status of the student model. MIM provides "instructional objects" to help students learn as they make progress. Teachers can also add their own instructional objects and attach them to the proficiency structure at the level of assessment claims. These instructional objects are presented to students depending on the assessment claims that the system holds about them.

Other tools for facilitating use of pieces of evidence include: tools for finding appropriate pieces of supporting evidence using external educational applications and linking them to evidence-based argument structures; tools for sharing supporting evidence with classmates or with the teacher in order to increase its strength; and alerting mechanisms that warn teachers, students, and parents of changes to argument structures. Future work includes developing some of these tools.

Using EI-OSM, teachers, students, and parents may become active participants in the assessment process. ECD-based learning environments that make use of EI-OSM can integrate evidence from various sources of evidence while providing valuable information right on time to help students learn.

Teachers play an important role in this approach. They are involved in many aspects of EI-OSM from defining acceptable pieces of supporting evidence to helping resolve potential conflicts. Since we want teachers to be able to use evidence of student performance maintained by assessment-based learning environments to help student learn, we interviewed eight Algebra teachers/assessment specialists regarding the features and feasibility of EI-OSM. We now present the results from our interviews with the eight teachers/assessment specialists.

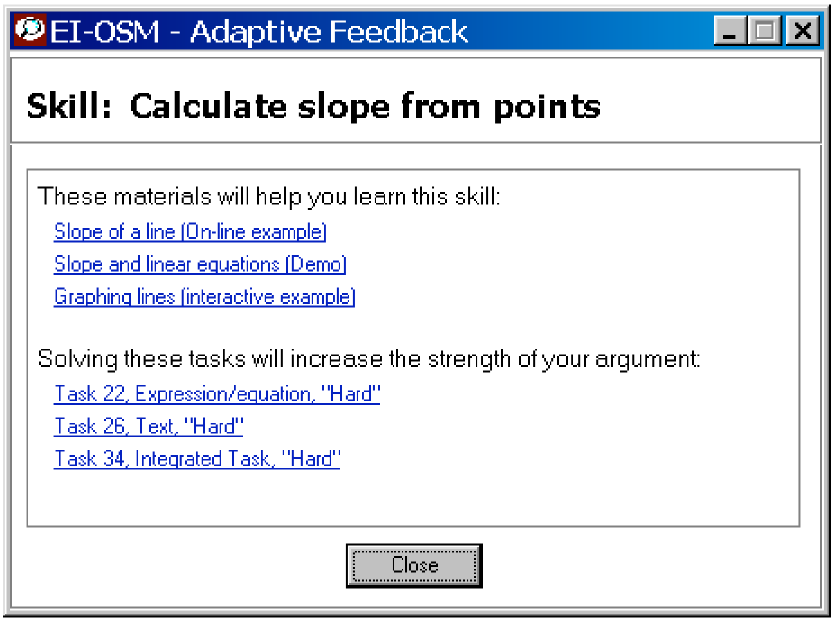


Fig.12. Providing adaptive feedback to students based on ECD model information.

EI-OSM EVALUATION

The main goal with this evaluation is to explore the feasibility of EI-OSM from the teachers' point of view. Main questions related to this goal are as follows: Do teachers understand proficiency maps? Can proficiency maps provide a useful overview of the student model and a starting point for exploring evidence-based argument structures? Do teachers understand evidence-based argument structures? Are evidence-based argument structures useful tools to be used in the classroom? How would teachers make use of evidence-based argument structures? What kinds of supporting evidence would teachers be willing to consider when interacting with EI-OSM (e.g. students' unsupported opinions, results from homework, and classroom assignments)? What degree of control over the system do teachers prefer? And what kinds of supporting mechanisms and tools should be available to successfully implement EI-OSM in real settings?

In order to answer these questions we interviewed eight Algebra teachers, who are also assessment specialists. This unique combination of skills made them ideal judges of this experimental approach. Interviews were conducted in small group settings (n = 1-3 teachers per setting with at least 2 interviewers per session), and lasted about 2 hours per session.

Three scenarios covering the main features of EI-OSM were implemented (see below). Each of the 3 scenarios was presented, one-by-one. After each scenario (or sometimes in the middle of a scenario), a series of questions were asked. Participants were then instructed to (a) write down their responses to the questions, and (b) discuss their responses/thoughts. The questionnaire used can be found in Appendix 1. The scenarios used in this study are as follows:

- *Scenario 1: Exploring assessment claims and supporting evidence.* Students have been using an assessment-based learning environment called Math Intervention Module (MIM) for a couple of weeks. MIM has gathered assessment information about students in the form of assessment claims and supporting evidence. In this scenario you will (a) interact with a map of proficiencies (*Scenario 1a*), and (b) explore some assessment claims and supporting evidence maintained by the system (*Scenario 1b)*.

- *Scenario 2: Exploring assessment claims and supporting evidence - Listening to Clarissa.* In this scenario you will use EI-OSM to explore some of the assessment claims and supporting evidence added by a student in response to assessment claims maintained by the system (*Scenario 2a*). In addition, we will talk about the role of teachers and students in EI-OSM (*Scenario 2b*).
- *Scenario 3: Assigning credibility and relevance values and providing adaptive feedback.* In this scenario you will use EI-OSM to add new pieces of supporting evidence and set initial credibility and relevance values. In addition, you will see how adaptive feedback is offered to students based on the student model.

Results

Results from our focused conversations with the eight Algebra teachers/assessment specialists were encouraging. Table 1 presents the general findings for each of the three scenarios described above. We provide more details on their perceptions below.

Scenario 1a - Understanding and using the Proficiency Map

In terms of the participants' general reactions to the "*Proficiency Map*" and associated claims, all participants understood the *Proficiency Map* as communicating a student's proficiency profile and relationships among proficiencies. All participants found it very useful as a representation of students' strengths and weaknesses. The characteristics of the map that were most appealing included: the color coding of the proficiencies (red, yellow, green), the relationships among the nodes, and the different possible paths they could create for instructional purposes (for different groups of students - which was not explicitly included in the design, but can be included in a subsequent version). When asked about which nodes to explore more fully first, most chose to explore the "weaker nodes" (red and yellow) first. Some thought it useful to examine branches of the map (e.g. important concepts like "understanding slope" while others thought it useful to examine clusters of similarly-colored nodes (for instructional purposes). The *Proficiency Map* was readily understood (typically a matter of minutes) by all - especially the color-coding and structure. Some took a little more time (again, in minutes) to understand directionality in the map (e.g. top nodes reflect direct and indirect - propagated from other nodes in the map - evidence).

Scenario 1b - Exploring MIM's assessment claims and supporting evidence

Concerning MIM's assessment claims and supporting evidence, all of the participants found the interface and associated information very useful for instructional purposes. However, some of them thought that more tasks should be included in the system as "evidence" to support the strength of a given claim in conjunction with an explanation of errors (e.g. diagnosis). For example, Expert 4 said "*More information is needed about the student's misconceptions. Information on prerequisites could be useful - at a finer grain size.*" Although this information was recently added to MIM based on error analysis, it was not included in this version of EI-OSM.

Table 1
Summary Findings from the EI-OSM Usability Study

Scenario	General Findings
Scenario 1a - *Understanding and using the Proficiency Map*	All 8 participants found the Proficiency Map to be very useful as a representation of students' strengths and weaknesses. The most appealing features of the map included: color coding of the proficiencies, clear relationships among the nodes, and obvious alternative instructional paths that could be pursued.
Scenario 1b - *Exploring MIM's assessment claims and supporting evidence*	Two broad functions of EI-OSM were noted by our participants: (a) to inform changes in classroom instructional strategies and topics, and (b) to help decide what to do next with individual (or small groups of) students. More specifically, they indicated that the information could be used to inform homework assignments, serve as the basis for further explanations to students, suggest alternative instructional approaches, recommend placement of students into small groups, and provide remediation as warranted.
Scenario 2a - *Exploring assessment claims and supporting evidence: Listening to Clarissa*	Our participants were split in their attitudes about how to handle unsupported claims by Clarissa. For example, and in relation to the scenario illustrated in Figure 9, when asked if they would accept Clarissa's alternative explanation, three said "no" and five said that they would - but assign it a low-strength value. But once supporting evidence was entered by Clarissa (shown in Figure 10), all 8 participants indicated that her alternative explanation was now acceptable. And in response to the question about whether they wanted to assign strength values or have the computer do so, half wanted to do it themselves, and the other half wanted the computer to do it.
Scenario 2b - *Exploring assessment claims and supporting evidence: Role of teachers and students*	All participants saw the utility of the approach for both teachers and students. They mentioned that teachers could use it to quickly identify gaps in understanding and focus on addressing those gaps (for individuals or small groups). They also mentioned that teachers could use it to chart instructional paths for the classroom, analyze new evidence, and focus teaching at higher levels with students (e.g. toward conceptual understanding). For students, they saw an important role of EI-OSM in helping students become more personally involved with (and accountable for) their own learning.
Scenario 3 - *Assigning credibility and relevance values and providing adaptive feedback*	In terms of assigning *credibility* values to evidence, our participants noted they would do so based on the source of evidence - such as if students worked alone or in a group, as well as the type of assignment. They would assign *relevance* values to evidence based on conceptual proximity to the topic. And with regard to doing it themselves or having the computer assign these values, they all were very keen to use tools to facilitate the assignment of values. Finally, when asked their preferences for use of numerical or descriptive labels, most wanted to see labels (e.g. "high") than numeric values for credibility and relevance.

When asked about what kinds of actions they would take based on the information offered by MIM, our respondents saw it could be used broadly in two ways: (a) to inform changes in classroom instructional strategies and topics, and (b) to help decide what to do next with individual students. They also indicated that the information could be used to inform homework assignments (class or individual level), serve as the basis for further explanations to the students (specifically) and suggest alternative instructional approaches to the topic (generally). Also, they wanted to use the information in the *Proficiency Map* to identify particular topics to review further and provide more thorough, annotated examples.

Regarding whether classroom teachers would like to use a tool EI-OSM, seven out of eight responded affirmatively. Some of the reasons mentioned included: "*It helps focus on aspects of instruction/learning to be emphasized*," "*Yes, especially if the program provides additional exercises at various levels ... and reports on common misconceptions*." The teacher who did not answer the question only mentioned that it would be useful only if adequate time was given to teachers to use the system.

In relation to teachers' willingness to use EI-OSM, all of them responded affirmatively. Key benefits included: helpful to teacher working with both groups and individuals, diagnosis of gaps in understanding, and special value for remedial math. Key context prerequisites for effective use include: adequate time, and sufficient content (diagnostic items, links to online and other instructional resources, adequate professional development, etc.)

When asked about what strategies they would use to make EI-OSM work in real settings, popular strategies included: (a) receive email alerts about students, especially groups (e.g. students holding common misconceptions), (b) involve tutors and teacher assistants in the process (i.e. tutors/teacher assistants can take care of individual cases), (c) ask students to contact the teacher during office hours when changes to the student model need to be approved, and (d) make use of progress reports - at group and/or individual levels.

Finally, participants perceived the main purposes of EI-OSM as: (a) using proficiency maps to create "learning paths," (b) using proficiency maps for diagnosis and outcomes (before/after instruction), (c) providing formative and summative assessments, (d) recommending placement into various small groups, and (e) providing remediation as needed. One of the participants mentioned that EI-OSM was ideal for remedial math. Some of the reasons cited included: modularity, tailored to the student, self-paced, and provision of guided assistance.

Scenario 2a - Exploring assessment claims and supporting evidence - Listening to Clarissa

When asked whether or not our participants would take into account Clarissa's alternative explanation given that no supporting evidence was available, three participants said "no," while five said that they would assign a low-strength value depending on the quality of Clarissa's explanation.

Next, when asked what kind of supporting evidence would be needed for them to take into account Clarissa's alternative explanation, participants mentioned additional problem solutions that would support Clarissa's claim and a detailed explanation of the solution. When asked how they would assign strength to Clarissa's claim once additional evidence had been provided, they indicated that they would increase or decrease the strength based on the quality of available evidence.

Finally, concerning whether or not they would take into account Clarissa's alternative explanation given that supporting evidence was available, all of them responded affirmatively. When asked how they would use available evidence to assign strength to Clarissa's claim, half of our participants wanted the system to automatically do it for them (4 out of 8). However, the other half wanted to have control over strength assignment via weights. In general, they agreed with the approach used in EI-OSM. It suggests the need to have user-configurable options for teachers to adapt the system to their preferences.

Scenario 2b - Exploring assessment claims and supporting evidence: Role of teachers and students

Participants noted that students play an active role in their own learning with EI-OSM. For example, one participant noted, "*Students take more responsibility for their learning, which is especially good at the college level*." Other participants supported this premise by mentioning that students can (a) add evidence/information/explanations to their student models, (b) keep track of their own learning, and (c) become aware of their evolving strengths and weaknesses. Teachers are assisted in quickly identifying gaps in understanding and focusing on addressing those gaps on an individual or group basis. Regarding the role of teachers, our participants mentioned that they would be able to: (a) design instructional paths for the classroom, (b) analyze new evidence, and (c) work at higher levels with students (e.g. toward conceptual understanding).

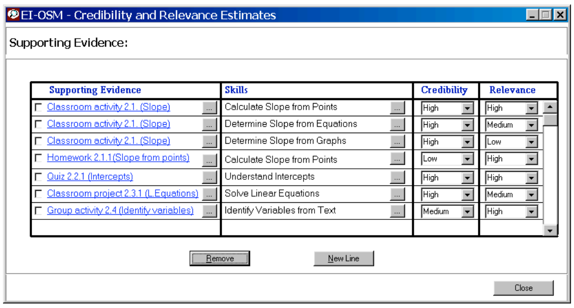
Scenario 3 - Assigning credibility and relevance values and providing adaptive feedback

When asked how they would assign credibility and relevance values, our participants suggested that they would assign credibility values based on the source of evidence - whether students worked alone or in a group, and the type of assignment (e.g. homework = low). Relevance values would be assigned based on the topic (i.e. the degree to which the evidence matches the skill). They were particularly interested in finding ways to facilitate this work (e.g. using a predefined list of credibility values per type of assignment).

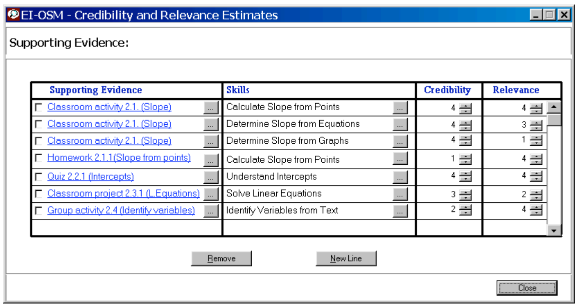
In response to our question about whether or not they would consider other sources of evidence (e.g. less credible pieces of supporting evidence such as unsupported student testimonies), 50% said "no" and 50% said "yes." When asked how they would combine credibility and relevance values into a single measure of strength, several participants made reference to using various weighting schemes.

Our participants were also asked to choose among three different graphical interfaces for assigning strength to pieces of supporting evidence (see Figure 13). The first graphical interface (Figure 13a) required the assignment of credibility and relevance values using labels (e.g. high, medium and low); the second one (Figure 13b) was similar to the first one but instead of labels, numerical values were used to represent levels; and the third one used numerical values to elicit strength values directly (Figure 13c). Most of our participants (5/8) preferred to separate credibility and relevance (as opposed to merging them into one single strength value. Among these five, four preferred labels over numerical values (i.e. Figure 13a). Participants welcomed the idea of using predefined packages of educational materials that provide information about relevance and credibility parameters automatically.

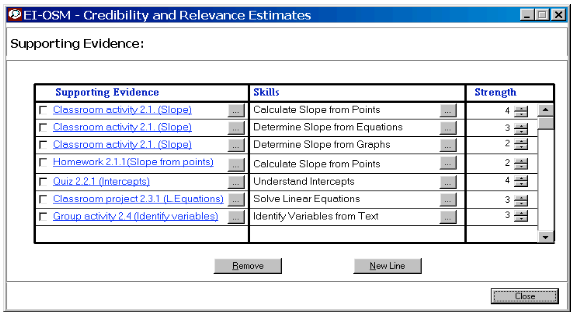
Finally, 7 out of 8 participants wanted the system to provide adaptive feedback (i.e. additional tasks and instructional objects) based on the student model. In addition, 6 out of 8 wanted to be able to add their own educational materials and tasks to the system.



a) Labels (Credibility and Relevance)



b) Numerical values (Credibility and Relevance)



c) Numerical values (Strength)

Fig.13. Assigning *Credibility*, *Relevance* and *Strength* values.

Discussion

Although teachers understood open evidence-based argument structures and acknowledged their potential for improving evidence-based communication with students and parents, they also warned us of several factors such as teachers' limited time and lack of resources that could jeopardize EI-OSM's successful implementation in the classroom. Teachers also suggested some interesting strategies that may be used to facilitate the implementation of EI-OSM in real settings (e.g. receiving email alerts, and involving tutors and teacher assistants in the process). Teachers seem to prefer a system that can operate autonomously (i.e. default algorithms should be in place to handle common cases) but also allow them to examine particular cases and override the system's behavior when considered necessary. The same line of reasoning applies to adding educational materials, tasks, and setting relevance and credibility values. That is, teachers would like to get predefined packages of educational materials and tasks (e.g. item banks) that can be automatically integrated into the system (i.e. links to KSAs and evidence parameters should be part of the predefined educational package) and at the same time would like to be able to change evidence parameters, remove tasks and/or educational materials, and add their own, again suggesting the need for configurable parameters/options.

Half of our participating teachers (4/8) were reluctant to consider unsupported assessment claims as evidence. They were aware of the need for appropriate evidence of student performance before making any instructional decisions. Some of them, however, were willing to accept these unsupported assessment claims when accompanied by a detailed explanation of how to solve previously failed tasks. Unsupported student opinions could be used to initiate a formative dialogue that could result in sophisticated evidence-based argument structures. On the other hand, students could use this feature to try to game the system. Moreover, EI-OSM uses different weights to assign strength values to conflicting assessment claims based on the evidence parameters of each piece of supporting evidence available. This evidence-based approach changes the nature of the problem from modeling how good the student is at self-assessing his or her KSAs, to what kinds of supporting evidence he/she can provide to strengthen the assessment claim. Assessing pieces of evidence and linking them to assessment claims is a problem we have investigated for a while using ECD. Modeling the evolution of assessment-based argument structures is a related problem that we can tackle by extending ECD principles and models.

While exploring the Proficiency Map, several teachers mentioned the idea of defining learning paths, grouping students, and using student model information to inform instructional strategies. Teachers were interested in using the Proficiency Map for formative purposes. Teachers also liked the idea of providing adaptive feedback, tasks and instructional objects based on the student model. EI-OSM can be used to link assessment information from formative and summative assessments. For example, assessment results and tasks from standardized tests can be used as part of an integrated assessment based learning environment, as presented in the AR approach. This is an interesting path that we are actively pursuing. An Assessment Based Learning Environment for English grammar (English ABLE) that makes use of enhanced TOEFL® tasks and was built based on ECD principles has already been used with real students (Zapata-Rivera et al., 2007).

EI-OSM is still in a development phase. We are in the process of adding new features to the system based on teachers' recommendations. Although based on initial feedback gathered we think that the approach presented in EI-OSM is feasible and promising, more studies are needed to clearly establish the benefits of this approach (e.g. studies involving students and parents). Some limitations of the current study include: (a) the teachers interacted with a version of the system that was limited to some hypothetical scenarios, (b) the number of teacher participants was relatively small, and (c) due to the informal nature of the interview some of the findings could not be easily quantified. We now present related research on evidence-based OSM.

RELATED WORK

Researchers have explored different approaches for involving the learner in diagnosis (e.g. Paiva & Self, 1995; Kay, 1999; Bull, 1997; Dimitrova, 2001; Zapata-Rivera, 2003). Each approach implements a different form of interaction with the student model and makes use of different modeling and reasoning techniques to represent students' and systems' beliefs and to resolve any conflicts that may arise from maintaining various views of the student model.

The student model server in TAGUS (Paiva & Self, 1995), for example, maintains the student's and the system's beliefs as Prolog clauses and uses belief revision techniques and a trust function to resolve possible conflicts.

e-KERMIT's (Hartley & Mitrovic, 2002) constraint-based student model is presented to the user in the form of domain categories that are selected based on their pedagogic importance. Students can also inspect a detailed hierarchical view of the student model. However, e-KERMIT does not support either direct modification of the contents of the student model by students or negotiation of the student model between teachers and students.

UM (Kay, 1999) maintains private and shared views of the user's and the system's beliefs. Each piece of evidence gets assigned a relative reliability value based on how the evidence was produced (e.g. direct, inferred), its source and date. Evidence parameters are used for reasoning about evidence, generating explanations, and resolving conflicts. The latter is implemented using *Resolver* modules that decide on the truth value of user model components by taking into account the type of each piece of evidence (i.e. given, observed, rules, and stereotype) and the time they were acquired. Thus, *Resolvers* act as plug-in components that can be used to model and reconcile a variety of conflicts involving various sources of evidence (e.g. student self-assessment of not knowing a concept vs. knowing it - Kay (1995)). It is at the end of the reconciling process that if two or more pieces of evidence happen to have the same reliability value, the most recent one is used by the *Resolver*. Kay and Lum (2005) describe the Scrutable Interface Viewer (SIV) that can integrate evidence from several sources using a weighting scheme. Weights are defined in relation to a "Standard Student" that performs at the highest level and completes all the assigned activities. Teaching assistants used SIV to explore various student models. Teaching assistants could compare students by looking at their student models. They also understood how various pieces of evidence were used to establish the student's knowledge levels. It is not clear, however, whether teaching assistants could add/remove pieces of evidence, influence the weighting scheme, or negotiate the contents of the student model with students. EI-OSM explores some of these issues. EI-OSM offers an approach aimed at facilitating teacher interaction with argument-based OSM. We have paid particular attention to how teachers perceive this environment and its potential to be integrated into their classrooms.

Mr. Collins (Bull & Pain, 1995) also supports student model negotiation between the student and the system by using a menu-based dialogue interaction. Negotiation mechanisms in Mr. Collins are developed based on Baker's negotiated tutoring approach (Baker 1990, 1994). The student and the system determine their confidence values (i.e. very sure, almost sure, unsure, very unsure) regarding the student's knowledge of a particular grammatical structure. The system uses information from the student's past five attempts at solving a task to determine its confidence value. If the student disagrees with the system's confidence value, he/she can try to change the system's confidence value ("justify myself") by answering an additional question. The student is offered a compromise (average of confidence values) without having to demonstrate his/her knowledge.

EI-OSM makes use of evidence-based argument structures to organize and integrate pieces of evidence. Alternative explanations or alternative claims gain strength as more evidence is available. These pieces of evidence can come from various sources. Teachers decide what pieces of evidence are acceptable. When the strength of an alternative claim surpasses the strength of the current assessment claim, it becomes the new assessment claim and its evidence is propagated throughout the Bayesian structure that supports the student model.

xOLM (Van Labeke, Brna & Morales, 2007) also implements an argument-based approach to interacting with open student models that is inspired by Toulmin's argument structure. However, xOLM's probabilistic student model includes cognitive, motivational and affective factors of the student and it is implemented using a variation of Dempster-Shafer Theory (Shafer, 1976). xOLM complements the graphical argument representation with natural language dialogue based on dialogue moves. This feature is used to facilitate human interaction with the model and at the same time serves as a log of the student's interactions with the system. In its current implementation, the student can challenge the system's claim by stating his/her proficiency and confidence levels, and providing an estimate of how strongly he/she believes the system to be right. EI-OSM also distinguishes from xOLM by the way evidence is handled (i.e. ECD and Schum's evidence based arguments) and by the central role that teachers play in this approach.

CONCLUSIONS AND FUTURE WORK

We have presented an evidence-based approach to interacting with open student models (EI-OSM). EI-OSM supports the refinement of assessment claims by using interactive, evidence-based argument structures that clearly show how evidence is used to support particular assessment claims. We have shown how evidence parameters (e.g. relevance and credibility) can be used to determine the "strength" of a particular assessment claim.

EI-OSM implements the Active Reports (AR) approach to opening ECD-based models, which makes evidence from existing summative and formative assessment available to teachers, students and parents as part of an assessment-based learning environment (MIM). EI-OSM also extends ECD to include assessment claims and supporting evidence originating from teachers, students and parents.

The evaluation results suggest that teachers understand the potential of EI-OSM for improving evidence-based communication and continuous, collaborative assessment. They valued evidence and were conscious of the need for valid and reliable sources of assessment evidence before making any claims regarding student performance and making any instructional decisions. Teachers wanted control over a system that can perform most tasks autonomously.

By creating, refining, and responding to evidence-based arguments, teachers, students, and parents are intended to become active participants within a new interactive assessment process in which evidentiary argument provides a basic language that unifies sources of assessment information. We think that evolution of evidence-based argument structures in EI-OSM offers an appropriate environment for developing tools aimed at enhancing student reflection and critical thinking skills. Open evidence-based argument structures serve as communication tools by clearly organizing and presenting assessment information used by teachers, students and parents.

EI-OSM represents an initial step toward implementing advanced interactive assessment environments that can be used in schools to enhance student learning and communication based on assessment information. Although our initial results with teachers are promising, we still need to study how students and parents react to this asymmetrical OSM approach in which teachers can override evidence from other sources. Thus we plan to conduct usability studies exploring how students and parents react to this instantiation of EI-OSM. We are also interested in exploring the effects of EI-OSM in relation to sustaining student motivation. Examining EI-OSM in a classroom environment will help us to determine the range of EI-OSM effects beyond the facilitation of evidence-based communication and the support of student learning (e.g. teachers' perceptions of students' capabilities may be enhanced). We are currently developing graphical interfaces for students and parents. New versions of the system will be used in future studies. We would like to continue exploring this line of research keeping in mind the various constraints inherent in real educational settings.

ACKNOWLEDGEMENTS

We would like to thank teachers/assessment specialists who participated in our study for their interesting comments and suggestions. We also would thank Irvin Katz, Lois Frankel, and three anonymous reviewers for providing valuable feedback on earlier versions of this paper.

REFERENCES

Baker, M. J. (1990). Negotiated Tutoring. An Approach to Interaction in Intelligent Tutoring Systems. Unpublished PhD thesis. Open University, Milton Keynes, UK.

Baker, M. J. (1994). A model for negotiation in teaching-learning dialogues. *Journal of Artificial Intelligence in Education*, 5(2), 299-254.

Bench-Capon, T. J. M. (1998). Specification and Implementation of Toulmin Dialogue Game. In J. C. Hage, T. J. M. Bench-Capon, A. W. Koers, C. N. J. de Vey Mestdagh & C. A. F. M. Grütters (Eds.) Jurix 1998: *Jurix*: *The Eleventh Conference.* (pp. 5-20). Nijmegen: Gerard Noodt Instituut.

Brna, P., Self, J., Bull, S., & Pain, H. (1999). Negotiated Collaborative Assessment through Collaborative Student Modelling. In *Proceedings of the Workshop on Open, Interactive and other Overt Approaches to Learner Modelling* (pp. 35-42). 9th International Conference on Artificial Intelligence in Education, Le Mans, France.

Bull, S. (1997). Collaborative student modelling in foreign language learning. Unpublished PhD Thesis. University of Edinburgh.

Bull, S. & Nghiem, T. (2002). Helping Learners to Understand Themselves with a Learner Model Open to Students, Peers, and Instructors. In P. Brna & V. Dimitrova (Eds). *Proceedings of Workshop on Individual and Group Modelling Methods that Help Learners Understand Themselves* (pp. 5-13). International Conference on Intelligent Tutoring Systems 2002, San Sebastian, Spain.

Bull, S., & Pain, H. (1995). Did I Say What I Think I Said, And Do You Agree With Me?: Inspecting and Questioning the Student Model. In J. Greer (Ed.) *Proceedings of the World Conference on Artificial Intelligence in Education* (pp. 501-508). Charlottesville, VA, USA: AACE.

Bull, S., Mangat, M., Mabbott, A., Abu Issa, A. S., & Marsh, J. (2005). Reactions to Inspectable Learner Models: Seven Year Olds to University Students. In J. Kay, A. Lum & J-D Zapata-Rivera (Eds.) *Proceedings of Workshop on Learner Modelling for Reflection, to Support Learner Control, Metacognition and Improved Communication between Teachers and Learners* (pp. 1-10). 10th International Conference on Artificial Intelligence in Education, Sydney, Australia.

Carbogim, D., Robertson, D., & Lee, J. (2000) Argument-based applications to knowledge engineering. *Knowledge Engineering Review*, 15(2), 119-149.

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L., & Ríos, A. (2004) SIETTE: A Web-Based Tool for Adaptive Testing. *International Journal of Artificial Intelligence in Education*, 14, 29-61.

Dimitrova, V. (2001). Interactive Open Learner Modelling. Unpublished PhD thesis, Computer Based Learning Unit, Leeds University.

Dimitrova, V. (2003). STyLE-OLM: Interactive Open Learner Modelling. *International Journal of Artificial Intelligence in Education,* 13(1), 35-78.

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics*, 33(1), 107-133.

Hartley, D., & Mitrovic, A. (2002). Supporting learning by opening the student model. In S. Cerri, G. Gouardères & F. Paraguaçu (Eds.) Intelligent Tutoring Systems: 6th International Conference, ITS 2002 (pp. 453-462). LCNS 2363. Berlin: Springer.

Jeong, A., & Juong, S. (2007). Scaffolding Collaborative Argumentation in Asynchronous Discussions with Message Constraints and Message Labels. *Computers & Education*, 48, 427-445.

Kay, J. (1995). The UM toolkit for cooperative user modelling. *User Modeling and User-Adapted Interaction*, 4, 149-196.

Kay, J. (1999). A Scrutable User Modelling Shell for User-Adapted Interaction. Ph.D. Thesis, Basser Department of Computer Science, University of Sydney, Sydney, Australia.

Kay, J., & Lum, A. (2005). Exploiting Readily Available Web Data for Reflective Student Models. In U. Hoppe, F. Verdejo & J. Kay (Eds.) Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies (pp. 338-345). Amsterdam: IOS Press.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30-43.

Kuntz, D., Fife, J., Shute, V. J., Graf, E. A., Supernavage, M., Marquez, E., et al. (2005). *MIM: Mathematics Intervention Module 1*. [Unpublished computer program/prototype]. Educational Testing Service, Princeton, NJ.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2000). Evidence-centered assessment design: A Submission for the NCME Award for Technical or Scientific Contributions to the Field of Educational Measurement. Retrieved December 10, 2006, from http://www.ncme.org/about/awards/mislevy.html

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective*, 1(1) 3-62.

Morales, R., Van Labeke, N., & Brna, P. (2006). Approximate Modelling of the Multi-dimensional Learner. In M. Ikeda, K. Ashley & T-W. Chan (Eds.) *Intelligent Tutoring Systems: 8th International Conference, ITS 2006* (pp. 555-564). Berlin Heidelberg: Springer-Verlag.

Paiva, A., & Self, J. (1995). TAGUS - a user and learner modelling workbench. *User Modeling and User-Adapted Interaction*, 4, 197-226.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., et al. (2005). The Assistment Project: Blending assessment and assisting. In C. K. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.) *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (pp. 555-562). Amsterdam: IOS Press.

Rouane K., Frasson C., & Kaltenbach, M. (2002). Advanced Annotation and External Representation in LKC System. *TICE 2002* (pp. 125-132).

Schum, D. (1994). *The Evidential Foundations of Probabilistic Reasoning*. New York: John Wiley & Sons.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Shute, V. J. (in press). Tensions, trends, tools, and technologies: Time for an educational sea change. To appear in C. Dwyer (Ed.) *The future of assessment: Shaping teaching and learning*. Mahwah, NJ: Erlbaum Associates.

Shute, V. J., Hansen, E., & Almond, R. (2007). Evaluation of ACED: The Impact of Feedback and Adaptivity on Learning. In R. Luckin, K. R. Koedinger & J. Greer (Eds.) *Artificial Intelligence in Education - Building Technology Rich Learning Contexts That Work* (pp. 230-237). Amsterdam: IOS Press.

Suthers, D., Weiner, A., Connelly, J., & Paolucci, M. (1995). Belvedere: Engaging Students in Critical Discussion of Science and Public Policy Issues. In J. Greer (Ed.) *Proceedings of the AIED'95 - 7th World Conference on Artificial Intelligence in Education* (pp. 266-273). Charlottesville VA: AACE.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nichols, R. Brennan & S. F. Chipman (Eds.) *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, United Kingdom: University Press.

Van Labeke, N., Brna, P., & Morales, R. (2007). Opening up the Interpretation Process in an Open Learner Model. *International Journal of Artificial Intelligence in Education*. 17, 305-338.

Verheij, B. (2001). Evaluating arguments based on Toulmin's scheme. Retrieved November 15, 2006, from http://www.ai.rug.nl/~verheij/publications/pdf/ossa2001.pdf

Zapata-Rivera, J.D. (2003) Learning Environments based on Inspectable Student Models. Ph.D. Thesis. Department of Computer Science. University of Saskatchewan. Saskatoon, Canada.

Zapata-Rivera, D., & Greer, J. (2002). Exploring Various Guidance Mechanisms to Support Interaction with Inspectable Learner Models. In S. Cerri, G. Gouardères & F. Paraguaçu (Eds.) *Intelligent Tutoring Systems: 6th International Conference, ITS 2002* (pp. 442-452). LCNS 2363. Berlin: Springer.

Zapata-Rivera, J.D., & Greer, J. (2003). Analysing Student Reflection in The Learning Game. In *Proceedings of the Workshop on Learner Modelling for Reflection* (Supplemental Proceedings Vol 5, pp. 288-298). International Conference on Artificial Intelligence in Education 2003, Sydney, Australia.

Zapata-Rivera, J.D., & Greer, J. (2004) Interacting with Bayesian Student Models. *International Journal of Artificial Intelligence in Education,* 14(2), 127-163.

Zapata-Rivera, D., Underwood, J.S., & Bauer, M. (2005). Advanced Reporting Systems in Assessment Environments. In J. Kay, A. Lum & J-D Zapata-Rivera (Eds.) *Proceedings of Workshop on Learner Modelling for Reflection, to Support Learner Control, Metacognition and Improved Communication between Teachers and Learners* (pp. 23-32). 10th International Conference on Artificial Intelligence in Education, Sydney, Australia.

Zapata-Rivera, D., VanWinkle, W., Shute, V. J., Underwood, J. S., & Bauer, M. I. (2007). English ABLE. In R. Luckin, K. R. Koedinger & J. Greer (Eds.) *Artificial Intelligence in Education - Building Technology Rich Learning Contexts That Work* (pp. 323-330). Amsterdam: IOS Press.

APPENDIX 1: EI-OSM QUESTIONNAIRE

Scenario 1: Exploring assessment claims and supporting evidence

1a. Exploring Clarissa's Proficiency Map (1)

Proficiency Map interface before selecting *CalculateSlopefromPoints* (see Figure 6)

1. What are the concepts that need more attention?
2. How would you use the proficiency map (e.g. explore red ones first, specific vs. general)?

Exploring Clarissa's Proficiency Map (2)

Proficiency Map interface after selecting *CalculateSlopefromPoints* (see Figure 6)

1. Would you choose a different proficiency? Why?
2. Would you rather explore several proficiencies before examining one in more detail?

1b. Assessment Claims and Supporting Evidence

Evidence Viewer interface (see Figures 7 and 8)

1. Is this interface useful? Is the information provided useful?
2. Based on these pieces of evidence what actions would you take? Some examples include:
   - o Talk to Clarissa about these tasks
   - o Choose different tasks and assign them to Clarissa (using the computer)
   - o Show Clarissa an annotated example
   - o Other : ___________________________
3. Do you think that teachers would like a tool like this, one that provides evidence and examples?
4. Do you think teachers would be willing to use this tool (e.g. teacher available time vs. helping students move ahead using evidence from EI-OSM)?
5. What strategies would you apply to make it work in real settings (e.g. classrooms)?
   - o Receive an email message (a warning) regarding Clarissa's progress
   - o Receive an email message when more than one student is experiencing similar problems
   - o Ask a teaching assistant to help you review particular cases
   - o Other:
6. How do you imagine teachers would use this tool (e.g. placement readiness, diagnostic purposes, summative purposes, practice keeping learning on track)?

Scenario 2: Exploring assessment claims and supporting evidence - Listening to Clarissa

Clarissa's Alternative Explanation

Evidence Viewer interface. No supporting evidence (see Figure 9)

1. Would you take into account Clarissa's alternative explanation (no supporting evidence)?
2. What kind of supporting evidence would you need to take Clarissa's claim into account?
3. How would you assign strength to Clarissa's claim?

Clarissa's Supporting Evidence

Evidence Viewer interface. Supporting evidence (see Figures 10)

1. Would you take into account Clarissa's new pieces of evidence?
2. How would you use this information to update Clarissa's assessment?
3. What do you consider is the role of the student in this assessment process?
4. What would be your role as a teacher?
5. How would you update the strength of Clarissa's claim?

Scenario 3: Assigning Credibility and Relevance Values and Providing Adaptive Feedback

Adding Supporting Evidence and Assigning Credibility and Relevance Values (1)

Credibility and Relevance interface (see Figure 11)

1. How would you assign credibility and relevance values (e.g. classroom activity 2.1 is highly credible but its relevance varies according to the skill used)?
2. Would you accept other sources of evidence (e.g. less credible pieces of supporting evidence such as unsupported student testimonies)?
3. How should relevance and credibility values be combined in a single strength value (one source and multiple sources of evidence)?

Credibility and Relevance interfaces (see Figures 13a, 13b, and 13c)

1. Which interface do you prefer? Why?

Providing Adaptive Feedback

Adaptive Feedback interface (see Figure 12)

1. Would you like to tell the system what to do in particular cases (adding rules to the system so it can help you assign activities automatically)?
2. What kind of feedback would you give to students who are using EI-OSM?